GIỚI THIỆU CÁC LOẠI ĐỒ THỊ THƯỜNG ÁP DỤNG TRONG THỐNG KÊ MÔ TẢ

(Tiếp theo kỳ trước)

Dominique Haughton - Nguyễn Phong

2. Số lượng cột trong đồ thị hình cột

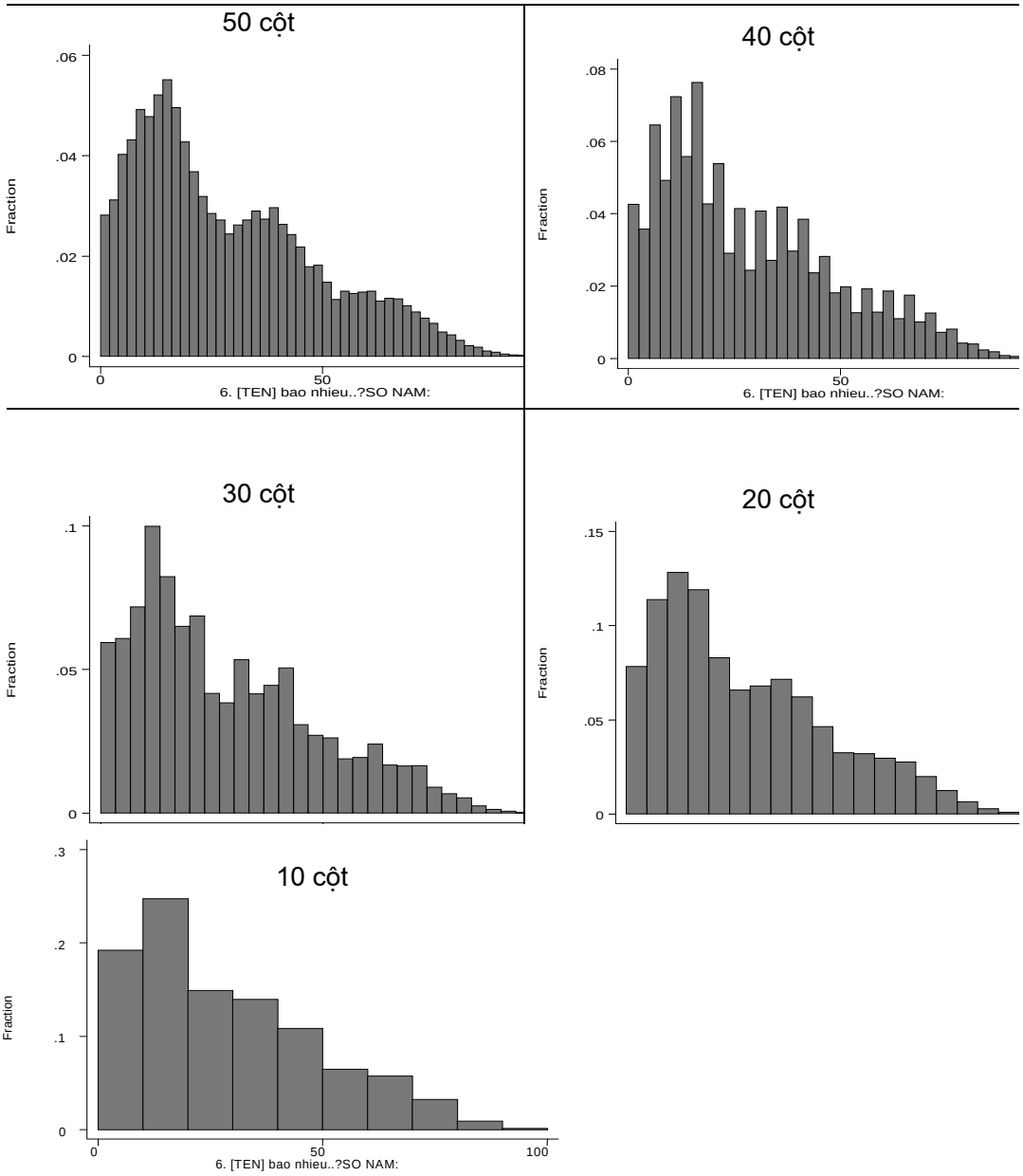
Đồ thị hình cột rất hữu ích trong việc đưa ra phân bố của một biến ngẫu nhiên, nhưng một khó khăn là phải chọn số cột sử dụng trong đồ thị. Rất nhiều phần mềm thống kê đặt ngầm định số cột nhưng số cột ngầm định này được đặt một cách khá chủ quan và thường có những hạn chế.

Dưới đây các tác giả đưa ra tác động của việc chọn số cột, từ 50 (số lớn nhất trong phần mềm Stata) đến 40, 30, 20 và 10 cột đến hình dạng của đồ thị hình cột. Biến được các tác giả sử dụng là độ tuổi của các cá nhân (được tính bằng năm) trong bộ số liệu VLSS98 (gồm 28.633 người có độ tuổi từ 0 đến 99).

Từ hình 5, chúng ta có thể thấy khi chọn số cột là 50 sẽ cho một bức tranh rõ ràng về ba thế hệ: trẻ, trung niên và già. Khi chọn số cột là 40 sẽ tạo ra một số đỉnh không thật, ví dụ các cột thứ 3 và thứ 5. Điều đó là do chiều rộng của mỗi cột là 2,475 (99/40), cột thứ hai nằm trong khoảng từ 2,475 đến 4,95 chỉ chứa những trẻ em trong độ tuổi từ 3 đến 4 tuổi trong khi cột thứ 3, nằm trong khoảng từ 4,95 đến 7,425, chứa nhiều trẻ em hơn vì nó chứa những trẻ em trong độ tuổi 5, 6 và 7 tuổi. Với số cột là 50, mỗi cột sẽ chứa 2 độ tuổi và do đó chúng ta sẽ không thấy các đỉnh được

tạo ra không thật như đã xuất hiện trên đồ thị có 40 cột. Với số cột là 10, bức tranh về 3 thế hệ không được nhìn thấy một cách rõ.

HÌNH 5: CÁC ĐỒ THỊ CỘT VỀ ĐỘ TUỔI CỦA CÁC CÁ NHÂN TRONG BỘ SỐ LIỆU VLSS 1998, VỚI SỐ CỘT LÀ 50, 40, 30, 20 VÀ 10



Câu hỏi đặt ra là liệu 3 nhóm có phải là con số phù hợp của các nhóm tuổi trong phân bố độ tuổi không. Tháp tuổi của Tổng điều tra dân số năm 1999 có thể khẳng định sự hiện diện của 3 nhóm tuổi này (xem Tổng cục Thống kê 2001). Tác giả Wand (1997) đưa ra một loạt các nguyên tắc "gán" để lựa chọn độ rộng (và

do vật là số) của các cột trong đồ thị hình cột. Sự biện giải cho các nguyên tắc này nằm trong một thực tế là đồ thị hình cột kết quả cung cấp một ước lượng mật độ gần nhất với mật độ thực xét theo sai số bình phương tích phân trung bình (*mean-integrated-square-error*). Việc tính toán các nguyên tắc này tương đối phức tạp nhưng chúng ta có thể có một ước lượng đầu tiên bằng việc sử dụng quy tắc "giai đoạn 0" (*zero-stage rule*). Quy tắc này xác định độ rộng của các cột bởi công thức:

$h_0 = 3{,}49\hat{}\, n^{-1/3}$, trong đó $\hat{\sigma} = \min\{S, IQR/1{,}349\}$, và *S* là độ lệch chuẩn của mẫu.

Quy tắc "giai đoạn 0" phù hợp với sự lựa chọn độ rộng của cột theo phương pháp "hệ thống chia độ thông thường" (*normal scale bin width selection*) được Scott (1979) đề xuất. So với các quy tắc khác do Wand đề xuất, quy tắc "giai đoạn 0" có thể ước lượng hơi quá lên độ rộng của cột nhưng vẫn là một sự hoàn thiện quan trọng so với các quy tắc được đặt ngầm định trong hầu hết các phần mềm thống kê trọn gói (Wand 1997, trang 63). Trong trường hợp của chúng ta, quy tắc "giai đoạn 0" tính ra độ rộng của cột là 2,314, do vậy số cột là khoảng 43 cột. Kết quả này biện giải thêm cho sự hiện diện của 3 nhóm tuổi trong bộ số liệu.

3. Mật độ Kernel

Mật độ kernel là dạng mở rộng của đồ thị cột, cho ta ước lượng về mật độ của một biến ngẫu nhiên.

Với bộ số liệu X_1, X_2, ..., X_n, đồ thị cột, hay một ước lượng mật độ đơn giản có thể được trình bày dưới dạng $\hat{f}(x) = (1/2hn)$[# X_1, X_2, ..., X_n] nằm trong khoảng (x-h,x+h). Hàm này sẽ bằng

$$\frac{1}{hn}\sum_{i=1}^{n}\left(\frac{x - X_i}{h}\right)$$

với w(x)=1/2 nếu $|x| < 1$, và w(x)=0 cho các giá trị khác của x. Nếu mở rộng hơn thì ước lượng mật độ Kernel được định nghĩa là hàm

$$f(x) = \frac{1}{hn}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

trong đó K là một hàm số và được gọi là k*ernel*. Tích phân $-\infty \div +\infty$ của K sẽ có giá trị bằng 1 nếu mật độ phân bố xác suất là đối xứng, ví dụ như mật độ phân bố chuẩn.

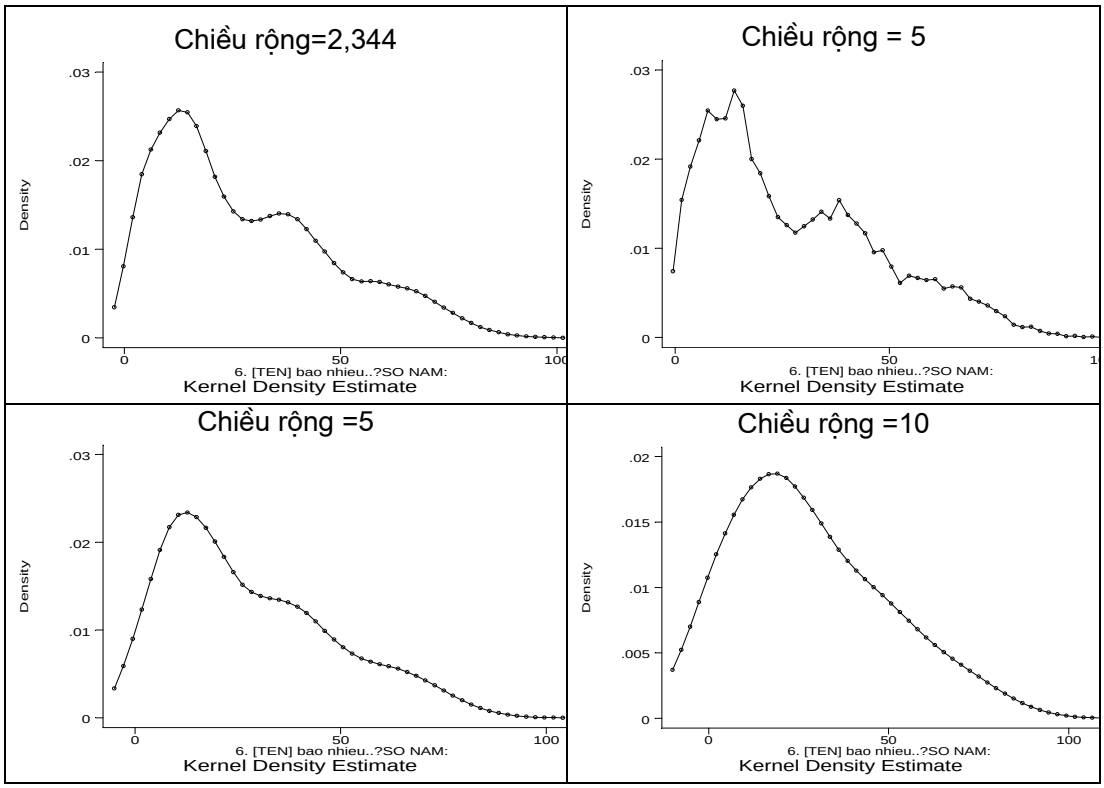
Lưu ý rằng cần phải lựa chọn được hàm kernel và chiều rộng h. Do vậy, về mặt này mật độ Kernel cũng có hạn chế tương tự như đồ thị cột. Chiều rộng h càng nhỏ thì càng nhiều chi tiết được thể hiện trên hình vẽ của mật độ kernel. Trong phần mềm Stata, lệnh *kdensity* ngầm định chiều rộng h là một hàm của cỡ mẫu, độ dàn trải và độ biến thiên của bộ số liệu. Hàm này được trình bày như sau: $h = 0.9\hat{\sigma}n\text{-}1/5$,với $\hat{\sigma} = \min\{S, IQR/1{,}349\}$ và S là độ lệch chuẩn của mẫu. Silveman (1986, trang 48) cho rằng sự lựa chọn h này đối với một hàm mật độ Kernel Gausian sẽ đưa ra sai số bình phương tích phân trung bình (từ mật độ thực tế đến mật độ ước lượng) nằm trong khoảng 10% của sự lựa chọn tối ưu cho một khoảng biến thiên lớn của mật độ thực tế. Mật độ Kernel Epanechnikov (được mô tả trong phần sau) sẽ cho một mật độ kernel gần giống với mật độ Kernel Gausian. Do vậy, sự lựa chọn h ngầm định này của Stata có nhiều khả năng phù hợp đối với hầu hết các trường hợp. Tuy nhiên, cũng cần lưu ý rằng với các phân phối xác suất có đuôi dài thì có

thể khó chọn được một chiều rộng với đủ các chi tiết cho phần chính của phân phối nhưng không quá nhiều chi tiết cho phần đuôi của phân phối.

Sự lựa chọn mật độ Kernel được khuyến nghị trong các tài liệu và được dùng ngầm định bởi lệnh kdensity trong Stata chính là mật độ Kernel Epanechnikov như đã giới thiệu ở trên. Mật độ kernel Epanechnikov là một hàm bậc hai lõm cắt trục hoành tại $\pm \sqrt{5}$. Đây là hàm mật độ kernel hữu hiệu nhất để giảm thiểu sai số khi dùng mật độ kernel để ước lượng mật độ thật (xem Silverman, 1986, Stata Corporation, 1999).

Trong Hình 6, chúng tôi trình bày 4 mật độ kernel đối với tuổi của từng cá nhân trong VLSS98. Mật độ kernel đầu tiên có chiều rộng của lớp là 2,344 (chiều rộng ngầm định) thể hiện rất rõ 3 nhóm tuổi: trẻ, trung niên và già. Chiều rộng 0,5 có thể cho nhiều đỉnh hơn mật độ thực tế nhưng vẫn thể hiện 3 nhóm tuổi. Mật độ kernel thứ ba với chiều rộng là 5 vẫn cho chúng ta hình dạng của 3 nhóm tuổi nhưng không rõ bằng mật độ với chiều rộng ngầm định. Và với chiều rộng bằng 10 thì ba nhóm này hầu như biến mất.

HÌNH 6. MẬT ĐỘ KERNEL ĐỐI VỚI TUỔI CỦA CÁC CÁ NHÂN TRONG VLSS98 VỚI CÁC CHIỀU RỘNG 2,344 (NGẦM ĐỊNH), 0,5 , 5 VÀ 10.

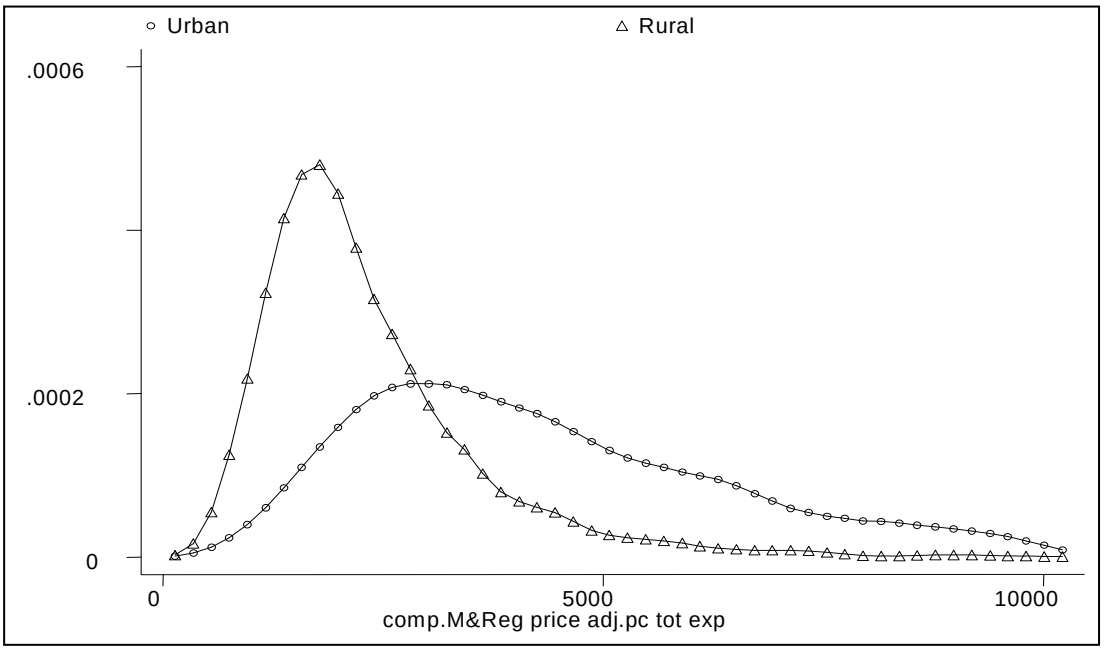


Đối với vấn đề chọn chiều rộng h, chúng tôi lưu ý một bài viết của tác giả Chiu (1991) đề cập đến một ước lượng giá trị chéo bình phương nhỏ nhất (*least squares cross validation estimator*) của chiều rộng tối ưu và phát hiện ra rằng

ước lượng này có khuynh hướng quá nhỏ và do đó có xu hướng thể hiện những chi tiết không có thật trong mật độ này (Chiu, 1991, trang 884). Tác giả Chiu đã khuyến nghị một kỹ thuật chọn chiều rộng ổn định, tốt hơn rất nhiều so với các ước lượng giá trị chéo, tuy nhiên rất phức tạp trong tính toán. Nên sử dụng kỹ thuật này khi cảm thấy nghi ngờ về hình dạng chi tiết của mật độ và khi hình dạng chi tiết này có tầm quan trọng lớn.

Mật độ kernel đặc biệt hữu ích trong việc so sánh hai hay nhiều nhóm: Trong Hình 7, chúng tôi vẽ mật độ kernel đối với chi tiêu thực tế bình quần đầu người của các hộ gia đình khu vực nông thôn và thành thị, sử dụng chiều rộng ngầm định trong Stata (424,45 đối với hộ gia đình thành thị và 161,60 đối với hộ gia đình nông thôn). Điều được thể hiện khá rõ trên đồ thị là phân bố của chi tiêu thực tế bình quân đầu người khu vực thành thị chuyển dịch sang phải và biến thiên hơn.

HÌNH 7: MẬT ĐỘ KERNEL CHI TIÊU THỰC TẾ BÌNH QUÂN ĐẦU NGƯỜI KHU VỰC THÀNH THỊ VÀ NÔNG THÔN ĐỐI VỚI VLSS 1998 VỚI CHIỀU RỘNG NGẦM ĐỊNH



(còn nữa)

Tài liệu tham khảo

Chiu, S.T. "Lựa chọn độ rộng cho ước lượng mật độ Kernel", *Biên niên sử thống kê*, 19, 1883-1905 (1991).

Hintze, J. and R.D. Nelson. "Đồ thị viôlông: Sự kết hợp giữa đồ thị hình hộp và đồ thị mật độ", *Tạp chí nhà thống kê Mỹ*, 52, 181-184 (1998).

Tổng cục Thống kê. *Niêm giám thống kê*. Hà nội: Nhà xuất bản thống kê (2001).

Rousseeuw, P., I. Ruts and J. Tukey. "Đồ thị hình túi: Đồ thị hình hộp biến thiên hai chiều", *Tạp chí nhà thống kê Mỹ* , 53, 382-387 (1999).

Scott, D.W. “Về đồ thị hình cột tối ưu và đồ thị hình cột dựa trên số liệu thực tế”, *Biometrika*, 66, 605-610 (1979).

Silverman, B. *Ước lượng mật độ cho số liệu thống kê và phân tích số liệu*. London: Chapman and Hall (1986).

Công ty Stata. "Lệnh kdensity", *Sách hướng dẫn Stata cho phiên bản 6*, tập 2, 144-151 (1999).

Tukey, J. *Phân tích số liệu thăm dò*. Reading, MA: Addison-Wesley (1977).

Wand, M.P. “Lựa chọn số cột cho đồ thị hình cột dựa trên các bộ số liệu thực tế”, *Tạp chí nhà Thống kê*, 51, 59-64 (1997).